Use of Multiple Assessments for Measuring Teacher Practice: An Initial Investigation for Educational Practice

Christopher M. Dudek

Ryan J. Kettler

Linda A. Reddy

Rutgers
The State University of New Jersey

Alexander Kurz
Arizona State University

Abstract

The present study examined the relationship between observer scores on three unique classroom observational assessments: the Danielson Framework for Teaching, the Classroom Strategies Assessment System, and the Instructional Learning Opportunities Guidance System. Each observational assessment measures teaching practices from a different theoretical perspective and methodological framework. Nine observers were trained to criterion for each measure and independently coded a series of ten classroom videos from teachers in Grades K to 12 using the FFT, CSAS, and MyiLOGS. Viewing order of the videos and measure completion for each observer was randomized to prevent presentation order bias. Pearson correlations were computed between measures. Findings indicated that the measures offer conceptually similar and dissimilar constructs, and yield complementary information for educator evaluation practices. Implications for research, practice, and policy are presented.

Keywords: classroom observation, multi-measure observation, teacher practice Running Head: Multi-Measure Classroom Observation

1. Introduction

In the recent era of education reform, teacher appraisal systems have become a significant focus worldwide. In March 2013, the International Summit on the Teaching Profession (ISTP), brought together an international audience and key stakeholders from multiple countries participating in the Programme for International Student Assessment (PISA; Organization for Economic Cooperation and Development [OECD], 2013) to explore how various countries address the issue of teacher appraisal. Themes of teacher quality, determinants for effectiveness, and the purpose of teacher evaluative functions were prominent among presentations (Walker, 2013; OECD, 2013). From the International Summit, the Teachers for the 21st Century report, an information source highlighting available research on effective teacher evaluation approaches and multiple examples of current teacher appraisal reform efforts world-wide was generated (OECD 2013). There is growing recognition that teachers significantly influence student learning and policy initiatives aimed at enhancing teacher quality and performance are likely to lead to positive student learning (Organization for Economic Cooperation and Development [OECD], 2005). For many countries, these teacher appraisal systems continue to be works in progress that require frequent revision as a result of new information on effective teaching, as well as incorporating stakeholders input, in particular teachers.

Similarly in the USA, teacher evaluation has been a topic at the forefront of education reforms the past decade. In response to federal legislative initiatives (e.g., No Child Left Behind, 2001; Elementary and Secondary Education Act, 2015) as well as seminal publications such as the "The Widget Effect" (Weisberg, Sexton, Mulhern, & Keeling 2009), teacher evaluation Running Head: Multi-Measure Classroom Observation models have made advancements toward multi-method measurement approaches (Mead, 2012).

As of 2014, 44 states in the US used teacher evaluation models that combine multiple measures of teacher practice (i.e., what teachers do in the classroom) and student achievement to evaluate teacher performance (National Council on Teacher Quality; NCTQ, 2000). From a US perspective, these changes are a positive sign of progress toward a multi-dimensional understanding of how to measure and improve teacher effectiveness.

1.1 Classroom Observations as Evaluation Components

In comparing international models to those in the US, classroom observations stand out as a common component in teacher appraisal processes. Generally, the intention of performing classroom observations is to capture key aspects of teaching and student-teacher interactions that are displayed during instructional periods. According to the OECD (2013), classroom observations are the most common source of evidence used in routine performance management of teachers in many countries (Isore, 2009; UNESCO 2007) and can also serve as evidence in special functions of appraisal systems. For example, observations of teacher performance influence the Korean performance-based incentive system, and in New Zealand they are used as part of the registration process for teachers (OECD, 2013).

In the new multi-method teacher evaluation models emerging across the US, classroom observations are typically the primary form of assessment used to measure teacher practice, and can carry as much if not more weight in evaluation schemes as student outcomes. In the US, most state education agencies (SEAs) require their school districts to use at least one classroom observational assessment to capture teacher practice, which is usually selected from a list of pre-approved measures. However, many school districts then rely on this one measure to observe and evaluate their teachers' practices. Internationally, this same issue appears in teacher appraisal Running Head: Multi-Measure Classroom Observation systems that are governed at the national level and rely on a singular framework for evaluating *all* teachers' performance. For example, Chile evaluates teachers using a national teacher appraisal system that contains four-domains and 20 criteria of assessment (Avalos &Assael, 2006) modeled after the Framework for Teaching (FFT; Danielson, 1996; 2007; 2013).

1.2 Disadvantages of Classroom Observational Assessments

Although a step forward in measuring teacher effectiveness, the "one size fits all" approach to measuring teacher practice through reliance on a single measure for classroom observations poses several questions related to validity and measurement bias. First, classroom observation instruments may vary in the constructs measured (Ko, Sammons, &Bakkum, 2013). Effective teaching strategies are often grouped under three big models related to direct instruction (explicit instruction), differentiated instruction, and constructivist methods. A classroom observational assessment that favors one of these models over others limits the range of available information on teaching practices to a narrow lens, limiting information that can be used to evaluate teacher performance and provide performance feedback for professional growth planning.

Second, classroom observational assessments can vary in their external validity in different educational and national contexts. Research has demonstrated that not all classroom observation instruments are suitable for use in all classroom contexts, with high poverty contexts and special education classrooms being prime examples (e.g., Brownell & Jones, 2014; Ko, Sammons, &Bakkum, 2013 Rowe, 2006). Students with learning disabilities benefit from instruction that is highly structured and contains ample opportunities for practice (Rowe, 2006). Socio-economically disadvantaged students benefit from instruction that is explicit and promotes basic skills (Scheerens, 1992; Muijs& Reynolds, 2005). Furthermore, studies contrasting Running Head: Multi-Measure Classroom Observation effective teachers in Hong Kong versus the United States suggests different cultural contexts can lead to an emphasis on different effective teaching practices and values (Jin&Cortazzi, 1998; Pratt et. al, 1999).

Third, there are differences in instrumentation that enable certain assessments to provide greater utility for summative assessments, whereas others may be better suited for formative purposes (Reddy, Fabiano, Dudek, & Hsu, 2013). Differences in instrumentation can also yield differences in reliability and validity. The Framework for Teaching (FFT; Danielson, 2013) is a widely used observational assessment that has been modified in multiple ways for evaluation internationally. As such, varying degrees of reliability and validity have been reported on the FFT, which can be attributed to the differences in measurement, training, implementation, and analysis each variation has used (Milaknowski, 2011). Lastly, there is a history of research documenting school-based observers, typically school administrators, as unreliable in evaluating teachers' classroom performance (e.g., Peterson, 1995; Porter, Youngs, &Odden, 2001; Toch& Rothman, 2008; Weisberg et al., 2009).

In sum, the valid use of teacher practice data from classroom observations to produce evaluation outcomes, professional development, or performance feedback, may be limited by the reliance on a single classroom observation measure.

1.3 A Multi-method Classroom Observation Approach

Akin to the multi-method approach used in measurement, the use of multiple classroom observational assessments offers one solution for addressing these issues. Numerous publications on teacher evaluation have highlighted the benefits of and advocated for multi-method approaches (e.g., Darling-Hammond, 2013; Goe, Bell, & Little, 2008; Holland, 2005; Kane, Kerr, Pianta, 2014). Best practices in evaluation from well-known organizations include the use Running Head: Multi-Measure Classroom Observation of multiple measures to inform decisions (AERA, APA, & NCME, 2014). Although key publications highlight the pitfalls of relying on a single source of information, current approaches to evaluating teachers' practices with classroom observations do not utilize a multi-method approach. Despite the large number of classroom observational assessments in use, limited research and educational practice information exists on the concurrent use of classroom observational assessments in schools (Kettler, Reddy et al., 2017). As Ko, Sammons, and Bakkum (2013) assert, "One of the most neglected areas in classroom observation research is using multiple instruments to examine the multidimensionality of teaching practices" (Ko, Sammons, &Bakkum, p. 29, 2013).

This article aims to address this gap by presenting a demonstration of the relations of three observational assessments that have been used in teacher evaluation. Specifically, we examined the relationship between observer scores on three unique classroom observational assessments: the FFT (Danielson, 2013), the Classroom Strategies Assessment System (CSAS; Reddy, & Dudek, 2014), and the Instructional Learning Opportunities Guidance System (MyiLOGS; Kurz et al., 2009). Each observational assessment measures teaching practices from a different theoretical perspective and methodological framework. The complimentary constructs captured in each approach offer opportunities to advance knowledge in teacher evaluation practices in schools.

1.3.1 Framework for Teaching

The FFT has existed for the past twenty years and is a well-known observation instrument world-wide (Danielson 1996; OECD, 2013). The FFT is standards-based instrument grounded in constructivist learning theory that examines teacher performance across four distinct domains (1) Planning and Preparation, (2) The Classroom Environment, (3) Instruction, and (4) Running Head: Multi-Measure Classroom ObservationProfessional Responsibilities (Danielson, 1996; 2013). It is designed to promote dialogue between evaluators and teachers by creating a shared understanding of effective teaching practices through the lens of higher order critical thinking conversations between teachers and students. Typically, the FFT produces scores along a four-category performance rubric (Unsatisfactory, Basic, Proficient, and Distinguished); school districts often assign a value of 1,2,3,4 to each category respectively, with higher scores being more desirable. The FFT is best characterized as on observational framework, and was not originally developed with modern test development theory (i.e., psychometrics) as a guiding principle. Teacher appraisal and evaluation systems around the world have been influenced by the FFT. As aforementioned, Chile uses a national appraisal system based on the four domains and criteria of the FFT (Avalos & Assael, 2006). England's earlier *Professional Standards for Teachers* (TDA, 2007) were inspired by the FFT. Quebec province in Canada, as well as multiple large school districts in the USA have also adopted custom versions of the FFT (Heneman et al., 2006; Isore, 2009; Milanowski, 2004).

1.3.2 Classroom Strategies Assessment System

In comparison, the Classroom Strategies Assessment System Observer Form (CSAS-O) is a recently developed multi-dimensional observational assessment that was designed using modern test development theory (Reddy, Fabiano, Dudek, & Hsu, 2013a). The CSAS-O was designed as a formative assessment to enhance teachers' usage of evidence-based classroom practices and like the FFT, aims to improve dialogue between evaluators and teachers about effective teaching practices. However, the CSAS-O includes multiple models of effective teaching (e.g., direct instruction, constructivist learning, adaptive instruction,) that emphasize a behaviorist model that focuses on teachers' behaviors (i.e., strategy usage) during instructional periods. The CSAS-O uses direct observation to inform frequency counts and Likert-type Running Head: Multi-Measure Classroom Observationbehavior rating scales that assess the extent to which teachers implement a specific set of evidence-based instructional *and* behavior management strategies during an observed lesson.

The CSAS-O produces several scores, including discrepancy scores [recommended frequency item ratings – observed frequency item ratings] which represent a need for change in classroom practices – the primary outcome that drives teacher professional development and can also be adapted to performance rubrics for evaluation purposes. Larger discrepancy scores indicated a greater need for change in classroom practices, thus smaller scores are desirable.

1.3.3 MyiLOGS

In contrast to the FFT and CSAS, MyiLOGS examines teachers' instruction as it relates to the construct of opportunity to learn (OTL). MyiLOGS is an online teacher log that allows teacher to report on their daily OTL provisions for their classes and individual students along three key dimensions of their enacted curriculum: time, content, and quality (e.g., Kurz, 2011; Kurz, Elliott, Kettler, &Yel, 2014). Each day, teachers record the amount of instructional time spent on state-specific academic standards as well as custom objectives. For each standard, teachers also record what cognitive processes students were expected to use, as well as various evidence-based instructional practices and grouping formats used during instruction. This information is used to calculate several OTL scores such as instructional time (IT), time on standards (TS), content coverage (CC), as well as scores indicating the emphasis of certain cognitive processes, instructional practices, and grouping formats. To estimate the accuracy of teacher self-report, MyiLOGS also provides an observation form that parallels the two self-report matrices used by teachers (i.e., cognitive processes by standards, grouping formats by instructional practices). For each observation, observers use an interval-based coding system to code the dominant cognitive process by standard intersect and Running Head: Multi-Measure Classroom Observation the dominant grouping format by instructional practice intersect. To this end, observers track each minute with a vibrating timer and make their matrix assignments during the 5-second vibration after each minute.

1.4 The Present Study

In sum, the FFT, CSAS and MyilOGS have been used in prior research, evaluation, and/or professional development. However, the three approaches have not been used concurrently to inform the appraisal process of instructional practices. Therefore, in the current article we examined the relationship between observers' scores on all three classroom observational assessments to offer a demonstration of the value in using multiple observational assessments in capturing unique and complimentary aspects of the instructional process. This study seeks to answer:

1. What are the within-measure relationships for scores of the FFT, CSAS-O, and MyiLOGS?
2. What are the between-measures relationships for scores of the FFT, CSAS-O, and MyiLOGS?

2. Method

2.1. Participants

2.1.1. Teacher Classroom Videos. In the current study, a series of 10 classroom videos were independently coded using the FFT, CSAS, and MyiLOGS. Each video represented a 30-minute segment of a reading, mathematics, or science lesson, and contained classrooms that spanned Kindergarten through twelfth grade. The video recordings transitioned from two perspectives: (1) a panoramic view of the whole classroom and (2) a view that focused on individual teacher-student interactions. The camera perspectives shifted whenever individual Running Head: Multi-Measure Classroom Observation interactions occurred. All videos were assessed for audio and visual quality prior to use in the current study.

2.1.2. Observers. Nine independent observers from a federally funded school reform grant, School System Improvement Project rated each of the videos. On average, observers were approximately 48 years old (SD =12.31; Range = 31 to 66) with majority of observers being female (66%). Observers identified their racial ethnicity as either Black/African American (66%) or Caucasian (33%). The majority of observers held a Master's degree (89%) and one observer possessed a doctoral degree (11%). On average, observers had 9.75 years of teaching experience (SD = 3.87 years; Range = 5 to 16) and for years of administrative experience, an average of 4.38 years (SD = 6.79 years; Range = 0 to 21). At the time of the study, the observers were acting as teacher and school administrator implementation coordinators for the SSI Project, and were responsible for facilitating project implementation efforts with partner schools.

2.2. Measures

2.2.1. Danielson Framework for Teaching (FFT; 2013). The FFT was developed from the Education Testing Service (ETS) PRAXIS III: Classroom Performance Assessments (Danielson, 1996; 2007; Dwyer, 1994). Grounded in constructivist learning theory, the FFT is a standards-based framework for evaluating teaching effectiveness. From this perspective, learners grow by developing their own understanding of concepts, hence the FFT directs observers toward students' actions and reactions as evidence for effective teaching. The FFT measures teacher performance across four distinct domains (1) Planning and Preparation, (2) The Classroom Environment, (3) Instruction, and (4) Professional Responsibilities. As displayed in Table 1, a total of 22 components (indicators of teacher performance) are nested within the four larger domains; the 22 components are composed of 76 smaller elements. During a classroom Running Head: Multi-Measure Classroom Observation observation, observers take notes relative to the components and elements within each domain. Observers then compare their notes to the descriptions and examples of these components in the framework's manual, and match their observations to a four-level performance rubric (Unsatisfactory, Basic, Proficient, and Distinguished). The FFT was updated in 2011 and 2013 to further enhance teacher evaluation practices and to include ideas from the Interstate New Teachers Assessment and Support Consortium (INTASC) standards.

Table 1. Components of the Danielson Framework for Teaching (FFT)

Domain	Components
1) Planning & Preparation1	1a. demonstrating knowledge of content and pedagogy
	1b. demonstrating knowledge of students
	1c. setting instructional outcomes
	1d. demonstrating knowledge of resources
	1e.designing coherent instruction
	1f. designing student assessments.
2) Classroom Environment	2a. creating an environment of respect and rapport
	2b. establishing a culture for learning
	2c. managing classroom procedures
	2d. managing student behavior
	2e. organizing physical space.
3) Instruction	3a. communicating with students
	3b. using questioning and discussion techniques
	3c. engaging students in learning
	3d. using assessment in instruction
	3e. demonstrating flexibility and responsiveness.
4) Professional Responsibilities1	4a. reflecting on teaching
	4b. maintaining accurate records
	4c. communicating with families
	4d. participating in the professional community
	4e. growing and developing professionally
	4f. showing professionalism.

The FFT has been implemented in teacher evaluation systems world-wide and numerous school districts are currently using the measure as intended or have made district-specific variations. Depending upon how it is implemented the FFT can provide scores at multiple levels. At the smallest unit of scoring, it is possible to match and assign each of the 22 components to one of four-level performance rubric categories. Traditionally, the four-level performance rubric categories are assigned to the domain level after examining all evidence collected for the components within each domain. An overall score is then created by looking across the four domains. In the current study, we used the 2013 version of the measure, which permits scoring rubric categories at the component level.

Because Domain 1: Planning and Preparation and Domain 4: Professional Responsibilities of the FFT rely on information beyond that which can be observed by video, the current study focused on Domain 2: Classroom Environment and Domain 3: Instruction. Both Domain 2 ($\alpha = .89$) and Domain 3 ($\alpha = .93$) were highly internally consistent in the current study. A small number of studies have been conducted on the reliability and validity of the FFT.

A large-scale assessment of the FFT's reliability occurred as part of the Measures of Effective Teaching (MET) study (Bill & Melinda Gates Foundation, 2012; 2013), however, reliability analyses only focused on inter-rater reliability. The study's authors found that 37% of variation Running Head: Multi-Measure Classroom Observationwas attributable to teacher differences, 43% of variation in FFT was attributed to other factors (i.e., rater, lessons, class section, time of year), and only a non- significant 10% attributed to lesson-to-lesson differences.

In a series of evaluation implementation studies of the Cincinnati public schools, Milanowski and colleagues (Milanowski , 2004; Milanowski, Kimball, & White, 2004) documented the inter-relationship of FFT domains for the 1996 version of the instrument. Correlations among the FFT's four domains were moderate with the exception of Planning and Professionalism (r = .75 for school year 2001 to 2002; r = .77 for school year 2001 to 2002) which was the highest, followed by the correlation between Classroom Management and Instruction domains correlations (r = .68 for school year 2001 to 2002; r = .61 for school year 2001 to 2002). Exploratory factor analyses did not yield evidence for the FFT four domains. Several studies have demonstrated evidence of the FFT's predictive validity with student achievement (Gallagher, 2004; Holtzapple, 2003; Kimball, et al., 2004; Milanowski, 2004). The strength of these associations have varied across studies, particularly in the areas of grade level and content, which may be due to differences in how the FFT was implemented (i.e., training, scoring, observation schedules) and the achievement tests used (Jones & Brownell, 2014).

2.2.2. CSAS Observer Form (CSAS-O). The CSAS-O is a multi-dimensional classroom assessment that measures teachers' use of evidence-based instruction and behavior management practices. The CSAS-O was designed as a formative assessment to facilitate teacher classroom practice improvements and can be used for summative evaluation, formative assessment, instructional coaching, pre-service mentoring, and research on teachers' classroom practices. The CSAS generates scores that (a) assess educators' use of empirically supported instructional and classroom behavioral management strategies, (b) generate professional development goals, (c) Running Head: Multi-Measure Classroom Observationmonitor teachers' progress toward goals, and (d) provide feedback for professional development (Reddy & Dudek, 2014). The constructs and items on the CSAS-O are based on models and strategies from over 50 years of effective teaching and behavioral management literatures (e.g., Alberto & Troutman, 2003; Brophy & Good, 1986; Hattie, Biggs, & Purdie, 1996; Horner, Sugai, Todd, & Lewis-Palmer, 2000; Kounin, 1970;Walberg; 1986; ; Stage & Quiroz, 1997; Walker, Ramsey, & Gresham, 2003). The CSAS-O includes three parts: (1) Strategy Counts encompass discrete counts of eight evidenced based teaching strategies, (2) Strategy Rating Scales consists of a 28 item Instructional Strategies (IS) rating scale and a 26 item Behavioral Management Strategies (BMS) rating scales and (3) the Classroom Checklist, which notes the presence or absence of key classroom structures and procedures. Observers complete the Strategy Counts during the observation period and take targeted notes relative to the scales and items of the Strategy Rating Scales and Classroom Checklist. Immediately after the observation, observers reflect on their notes and complete the Strategy Rating Scales: Instructional and Behavior Management Strategies (see Table 2 for definitions). The Classroom Checklist is completed before the observer leaves the room. In the current study, only scores from the Strategy Rating Scales were utilized.

Table 2. Descriptions of the CSAS Part 2 Rating Scales

Scale Name	Definition
Instructional Strategies Total (IS)	The Total IS Scale Reflects the overall use of Instructional Methods and Academic Monitoring/Feedback
Adaptive Instruction (AI)	Strategies teachers use to respond to their students' learning needs while teaching. These practices reflect teacher flexibility and responsiveness to students' needs, as well as methods of differentiated instruction.
Student-Directed Instruction (SDI)	Strategies teachers use to actively engage students in the learning process. These practices encompass constructivist and hands-on instructional techniques, linking lesson content to prior learning, personal experiences, and cooperative learning.
Direct Instruction (DI)	Strategies teachers use to deliver academic content or convey information to students. These practices include direct instruction techniques, modeling, identifying, and summarizing.
Promotes Students' Thinking (PST)	Strategies teachers use to activate students' thinking about the lesson material. These practices assess teachers' efforts to get their students to think about their thinking process (i.e, open-ended, what, how, why).
Academic Performance Feedback (APF)	Strategies teachers use to provide specific feedback to their students on their understanding of the material. These practices assess teachers' efforts to explain what is correct or incorrect with student academic performance. These practices also measure teachers' efforts to reinforce (i.e, praise) students learning.
Behavioral Management Strategies Total (BMS)	The Total BMS scale reflects the overall use of Prevention Methods and Behavior Feedback
Proactive Methods (PM)	Verbal and nonverbal strategies teachers use to prevent student disengagement, and problem behaviors from occurring in classroom. These practices assess how teachers create a positive classroom environment.
Directives (PM)	Strategies teachers use for issuing directions or instructions to students and behavioral expectations in the classroom.
Praise (P)	Verbal and nonverbal strategies teachers use to positively reinforce specific appropriate behaviors in the classroom. These practices assess how teachers respond to positive behavior in the classroom.
Corrective Feedback (CF)	Verbal and nonverbal strategies teachers use to correct students' inappropriate behavior. These practices asses how teachers respond to negative behavior in the classroom.

The Instructional Strategies Rating (IS) scale contains 28 items covering the areas of Adaptive Instruction, Student-Directed Instruction, Direct Instruction, Promotes Students' Thinking, and Academic Performance Feedback. The Behavioral Management Rating (BMS) scale contains 26 items covering the areas of Proactive Methods, Directives, Behavior Praise, and Behavior Corrective Feedback. To complete the Strategy Rating Scales, observers first rate how often (Observed Frequency rating) teachers used each of the strategies on a seven-point scale (1 = *Not Used*, 4 = *Sometimes Used*, 7 = *Always Used*). Observers then rate how often Running Head: Multi-Measure Classroom Observation teachers should have used those strategies (Recommended Frequency rating) in the observed lesson using the same seven-point scale. Ratings for recommended frequency are made based on three considerations: (1) the instructional objectives for the observed lesson, (2) research-based guidelines for use of specific strategies, and (3) observed learning and behavioral outcomes for teachers and students during the classroom observation. Discrepancy scores are calculated for each item by subtracting the observed frequency rating from the recommended frequency rating and taking the absolute value of the difference [Σ |Recommended Frequency – Observed Frequency|]. The difference between the two scores represents a need for change in a particular strategy, with larger differences indicating a greater need for change. Scale scores are then created by adding the corresponding items together.

The CSAS-O has demonstrated good content and construct validity, as well as reliability indices (Reddy, Fabiano, Dudek, & Hsu, 2013a). In the current study, the IS (α = .92) and BMS (α = .91) were both internally consistent. In previous research, the instrument has demonstrated good inter-rater (r = .94 for Strategy Counts; r = .80 for IS rating scale; r = .72 for BMS rating scale) and test-retest reliability (r = .70 for Strategy Counts; r = .86 for IS rating scale; r = .80 for BMS rating scale). The IS and BMS Scales are theoretically and factor analytically derived with confirmatory factor indices evidencing good fit, as well as good internal consistency (r = .91 for IS rating scale; r = .92 for BMS rating scale). The CSAS-O items have also demonstrated freedom from item bias (Steele, House, &Kerins, 1971) on key variables that have been traditionally associated with differences in teaching quality (e.g., age, years of experience, degree). The CSAS-O has been found to have good convergent and divergent validity with classroom observational assessments such as the Classroom Assessment Scoring System (CLASS; Pianta, LeParo, & Hamre, 2008; Reddy, Fabiano, & Dudek, 2013) and student ratings Running Head: Multi-Measure Classroom Observation of the classroom environment with the Responsive Environment Assessment for Classroom Teaching (REACT: Nelson, Reddy, Dudek, & Lekwa, 2017) .

Also, the CSAS has been found to evidence predictive validity of students' proficiency status on state-wide assessments (Reddy, Fabiano, Dudek, & Hsu, 2013b), student academic engagement as measured by the Cooperative Learning Observational Code for Kids (CLOCK, Volpe & DiPerna, 2010; Lekwa, Reddy, & Shernoff, 2017) and student growth as measured by the Measures of Academic Progress (MAP; Lekwa, Reddy, Dudek, & Hua, 2017).

2.2.3. Instructional Learning Opportunities Guidance System (MyiLOGS). In the current study, the observer form of MyiLOGS, which is a paper-and-pencil form that parallels the teacher self-report matrices of the online system, was used to measure teachers' practices. For purposes of this study, observers completed the first (cognitive processes by standards) matrix following the 1-minute interval without differentiating by multiple standards. Given that the standards covered during each video were not available, observers identified the dominant cognitive process only (rather than also assigning each cognitive process to a list of available standards). The five cognitive processes contributing to the CP score are: Attend, Remember, Understand/Apply, Analyze/Evaluate, and Create, all of which are adapted from Bloom's taxonomy and supported by Webb's Depth of Knowledge. For the second matrix (grouping formats by instructional practices), observers recorded the dominant instructional practice and the grouping format used to implement the respective practice. The nine instructional practices contributing to the IP score are: Provided Direct Instruction, Provided Visual Representations, Asked Questions, Elicited Think Aloud, Provided Guided Feedback, Provided Reinforcement, Assessed Student Knowledge, Used Independent Practice, and Other Instructional Practices. Running Head: Multi-Measure Classroom Observation The logged and observed information from these two matrices can be used to calculate several OTL scores. All three scores represent the percentage of time spent in one of two categories (higher-order cognitive processes vs. lower-order cognitive processes, evidence-based instructional practices vs. generic instructional practices, small/individual grouping formats vs. whole class grouping formats; see Table 3).

Table 3. ***Definitions of MyiLOGS indices and components***

Name	Definition
Student Cognitive Processes (CP)	Amount of instructional time dedicated to higher order student cognitive processes[3]
Attend	Orient toward instructional task and related instructions
Remember	Retrieve relevant knowledge from long-term memory
Understand/Apply[1]	Construct meaning from instructional messages/carry out or use a procedure in a given situation
Analyze/Evaluate[1]	Break materials into its constituent parts and determine how the parts relate/make judgements based on criteria and standards
Create1	Put elements together to form a coherent whole or a new structure
Teacher Instructional Practices (IP)	Amount of instructional time dedicated to empirically supported or evidenced-based practices.[3]
Provided direct instruction[2]	Teacher presents issue, discusses or models a solution approach, and engages students with approach in similar context
Provided visual representations[2]	Teachers uses visual representations to organize information, communicate attributes, and explain relationships
Asked questions[2]	Teacher asks questions to engage students and focus attention on important information.
Elicited think aloud[2]	Teacher prompts students to think aloud about their approach to solving a problem
Used independent practice[2]	Teacher allows students to work independently to develop and refine knowledge and skills
Provided guided feedback[2]	Teacher provides feedback to students on work quality, missing elements, and observed strengths or work performance
Provided reinforcement[2]	Teacher provides reinforcement contingent on previously established expectations for effort and/or work performance
Assessed student knowledge	Teacher uses quizzes, tests, student products, or other forms of assessment to determine student knowledge
Other instructional practices	Any other instructional practices not captured by the aforementioned key instructional practices
Grouping Formats (GF)	Amount of instructional time dedicated to individual &/or small group instruction.[34]

Note. 1 – Indicates higher order critical thinking processes for students; 2 – Indicates evidence based teaching practice; 3 – The CP, IP, and GF scores were measured as a percentage of time instead of total discrete minutes. All other scores are measured as a discrete number of minutes. 4 – In the current study, Grouping Formats were not included in the analyses.

The first seven instructional practices have received substantial empirical support from research syntheses and meta-analyses (e.g., Brophy & Good, 1986; Gersten et al., 2009; Marzano, 2000; Vaughn, Gersten, & Chard, 2000; Walberg, 1986). Other Instructional Practices represents a generic category for teachers to report on practices The grouping formats used to determine the GF score are defined as (a) Individual: instruction focused on individuals working on different tasks, (b) Small Group: instruction focused on a small group working on different tasks, and (c) Whole Class: instruction focused on the whole class working on the same task. For purpose of this study, we focused on the quality-related OTL scores of CP, IP, and GF.

The psychometric properties of MyiLOGS have been documented by earlier research (e.g, Kurz, Elliott, Lemons, et al., 2014) and the evidence (i.e., response processes, internal structure) for the validity of inferences is acceptable. Internal structure validity analyses of MyiLOGS indicated that the various indices measured relatively independent constructs, with no pair of the five sharing a correlation greater than r = .38 (Kurz, Elliott, Kettler, &Yel, 2014). Construct validity was evidenced by high agreement between MyiLOGS scores and MyiOBS

observations (percent agreement = 77%). Agreements between log data from teachers and independent observers were comparable to agreements reported in similar studies. Based on data from these studies, the evidence indicates that (a) MyiLOGS has high usability, (b) its quarterly summary scores are relatively consistent across time, and (c) summary scores based on random Running Head: Multi-Measure Classroom Observation samples of 30 calendar days and ten detail days can provide reliable estimates of teachers' respective yearly summary scores.

2.3. Procedure.

2.3.1. Training. Observers received training on all three assessments prior to the start of the study and passed all required certification tests for each measure. For the FFT, observers participated in the online certification training product sponsored by Teachscape. On average the online course takes approximately 36 hours to complete, and exposes observers to the constructs supporting the FFT's Domain 2: Classroom Environment and Domain 3: Instruction. Observers engage in several practice videos and knowledge tests during the course of training. At the end of training, observers are required to take a video certification that is composed of three videos, each 30 minutes in length. Videos are randomly selected for each observer and the minimum score required to pass is 70%.

For the CSAS-O training consisted of a three-day group training led by the authors of the CSAS, which focused on CSAS theory, definitions, and administration and score interpretation (including knowledge tests and video coding practice). Observers were then required to engage in several practices videos until they were comfortable using the CSAS-O. To achieve reliability on the CSAS-O, the observers were required to pass a video coding criterion test that consisted of five classroom videos, each 15 minutes in length. Observer scores were compared to the master codes for each video and required to pass with a minimum threshold of 70%.

For MyiLOGS, observers were first trained on the online teacher self-report tool, which requires passing a knowledge test at the highly qualified criterion (i.e,, score greater than 80%). Observers were subsequently trained on the MyiLOGS observer form using classroom videos. Running Head: Multi-Measure Classroom Observation Following model and guided practice video sessions, all participants independently coded two 20-minute video segments and passed the criterion of 80% agreement on two videos.

2.3.2. Video coding. Each independent observer completed all three instruments on the ten classroom videos according to a randomized schedule that varied the order of observational instruments and video. For example, the presentation order and video for an observer may have followed the order of: (1) Video #1 – FFT, Video #4 – MyiLOGS, (3) Video #2 – FFT, (4) Video #9 – CSAS, (5) Video #7 – CSAS, etc. The randomized order was created using a random number generator. Thus each video was assessed separately using one observational instrument at a time to prevent presentation order interference.

3.0 Results

Scores on the CSAS-O, FFT, and MyiLOGS were comparable to those observed in other research. Mean CSAS-O Strategy Count scores and Strategy Rating scale discrepancy scores for both IS and BMS scales were within ¼ SD of scores from previous studies (Reddy et al., 2013b). Mean ratings for both FFT domains were closer to 3.0 (Proficient) than they were to 2.0 (Basic). The mean MyiLOGS percentage of time spent in higher order CP (40%) was lower than in previous research (74%; Kurz et al. 2014). The percentage of time spent using evidence-based IP (57%) and small-group GF was comparable to previous research. Table 4 depicts means and standard deviations for scores from the three observational measures.

Table 4. *Means and Standard Deviations of Scores from Observational Measures*

	Scale [a]	n	Mean	SD
CSAS-O	Instructional Strategies	87	23	16
	Behavioral Management Strategies	87	17	14
FFT	Domain 2: Classroom Environment	85	2.80	.74
	Domain 3: Instruction	85	2.66	.84
MyiLOGS	Cognitive Processes	90	40%	24%
	Instructional Practices	90	57%	24%
	Grouping Formats	90	27%	24%

CSAS-O = Classroom Strategies Assessment System – Observer form; FFT = Danielson Framework for Teaching; *Significant at $\alpha < .05$, 1-tailed.

3.1 Within Measures Relationships

Pearson correlations were calculated between scores within measures for the CSAS-O, FFT, and MyiLOGS. Cohen's (1988) recommendations for describing correlational magnitude were used; values below .10 are non-substantial; values between .10 and .30 are small; values between .30 and .50 are moderate, and values greater than .50 are large. Table 5 depicts Running Head: Multi-Measure Classroom Observation correlations for scores within and between the three measures. For the CSAS-O, the correlation between discrepancy scores for the IS and the BMS (r = .69) was in the large range. For the FFT, the correlation between Domain 2: Classroom Environment and Domain 3: Instruction (r = .79) was in the very large range. For MyiLOGS, correlations between CP and the other two scores were in the medium range. The correlation between IP and GF was in the small range (r = .17).

Table 5. Correlations between CSAS-O Strategy Rating Scales, the FFT, and MyiLOGS

		CSAS-O		FFT		MyiLOGS		
		Instructional Strategies	Behavioral Management Strategies	Classroom Environment	Instruction	Cognitive Processes	Grouping Formats	Instructional Practices
CSAS-O	Instructional Strategies	-						
	Behavioral Management Strategies	.69*	-					
FFT	Classroom Environment	-.38*	-.43*	-				
	Instruction	-.34*	-.38*	.79*	-			
MyiLOGS	Cognitive Processes	-.36*	-.49*	.46*	.34*	-		
	Grouping Formats	-.13	-.20*	.19*	.04	.40*	-	
	Instructional Practices	-.11	-.30*	.15	.05	.35*	.17	-

Note. CSAS-O = Classroom Strategies Assessment System – Observer form; FFT = Danielson Framework for Teaching; *Significant at $\alpha < .05$, 1-tailed.

3.2 Between Measures Relationships

Based on the six scores of interest generated by the three measures, 14 correlation coefficients were calculated between scores yielded by two different measures. Generally, relationships between CSAS-O and the other two measures were negative in direction, whereas relationships between FFT and MyiLOGS were in the positive direction. Of these coefficients, nine had magnitudes in the medium range and six had magnitudes in the small range. All correlations between FFT and CSAS-O, as well as all correlations involving CP from MyiLOGS, were in the medium range. IP and GF from MyiLOGS diverged more from the other measures, sharing correlations with magnitudes of less than or equal to .30 with all scores.

4.0 Discussion

4.1 Reliability of FFT, CSAS-0, and MyiLOGS

Findings from the current study indicate the measures are internally consistent and yield scores consistent with previous research. The FFT scores, based on components of each domain, were found to be internally consistent in the current study, with alpha estimates considerably higher than those found in previous research (Milaknowski , 2004; Milaknowski, Kimball, & White, 2004). As aforementioned, the FFT permits modifications to the framework and numerous adaptations and modifications of the framework exist internationally (Avalos &Assael, 2006; Milanowski, 2011; Isore, 2009). In a review of observation measures, Goe, Bell Running Head: Multi-Measure Classroom Observation and Little (2008) indicate that for the FFT there is "wide variation in rater training, rater's relationship with the teacher, the degree of adherence to Danielson's recommendations for use, the use of scores, and the number of observations conducted for each teacher" (p.23). The authors further stipulated that research on the comparative performance between modified versions of the FFT and models adhering to Danielson's recommendations are inconclusive. Along those lines, more recent research emerging from the large-scale Measures of Effective Teaching (MET) study, indicated that a modified version of the FFT was only modestly correlated with students' academic achievement (Kane et al., 2013), a finding similar to previous FFT research.

Although different than Danielson's recommendations, the FFT scoring approach used in the current study is similar to scoring methodology and scale score calculations used in traditional psychometric development approaches (Kettler & Reddy, in press). Based on the current study's results, it appears that using a more rigorous scoring approach for determining the FFT domain scores enhances the reliability of the measure. School districts' implementing the FFT, even in modified formats and adaptations, should consider scoring approaches that enhance the accuracy and reliability of the instrument.

The CSAS-O scores yielded values congruent with previous research and the Strategy Rating scales were also similarly found to be internally consistent at levels congruent with previous research. Subsequently, the current study's results contribute to the evidence base supporting the reliability of the CSAS-O (Reddy, Fabian, Dudek, & Hsu, 2013a). The MyiLOGS scores in the current study varied in their congruence with previous research, most notably with the CP index being much lower. However, this could be due to the time-sampling method of the MyiLOGS observer form, which has been shown to have limitations due to the smaller interval Running Head: Multi-Measure Classroom Observation of time captured by observers compared to teachers' self-report logs, which can encompass the entire school day (Kurz, Elliott, Lemons, et al., 2014).

4.2 Relationships between Measures

We found that the FFT's domains demonstrated significant negative correlations in the medium range with the CSAS-O Strategy Rating scale discrepancy scores. As FFT scores decreased, suggesting lower quality classroom environment and instruction, the CSAS-O discrepancy scores increased, indicating a greater need for change in classroom practices. This finding reflects the realistic outcome of a poor classroom observation and the subsequent feedback delivered to the teacher. Typically, teachers found to be less effective via observations with the FFT would receive feedback indicating they need to make significant changes in their practices. Conversely, the current study's results also suggest that teachers found to be more effective on the FFT, would receive feedback indicating no change or minimal changes to their classroom practices are needed. Previous concurrent validity studies that have correlated observational assessments with positive scoring schemes (e.g., CLASS, Pianta, LeParo& Hamre, 2008) to the CSAS-O Strategy Rating scale discrepancy scores have found a similar negative correlation trend between constructs (Reddy, Fabiano, & Dudek, 2013; Reddy, Dudek, Rualo, & Fabiano, 2016) and the current findings add to the validity base for the CSAS-O's Strategy Rating scale discrepancy scores.

The pattern of relationships between the FFT scores and the MyiLOGS scores reflected behavior management (Domain 2: Classroom Environment) sharing significant correlations with CP and GF, while the score reflecting instruction (Domain 3: Instruction) only shared a significant correlation with CP. The direction of these relationships match the theoretical content-validity alignment direction and provide additional validation evidence for both Running Head: Multi-Measure Classroom Observation measures. When there is high quality instruction and a positive classroom environment, the presence of teaching strategies focused on higher order critical thinking processes increases. This relationship between the two measures is to be expected given the CP index and scores focus on students' cognitive processes, with a preference for higher order critical thinking processes, and the FFT as a whole, favors dialogue between teachers and students that occurs at higher order critical thinking and metacognitive levels.

The FFT's domain scores and the MyiLOGS indices of GF and IP, did not evidence significant relationships. This may be due to a theoretical content-validity mismatch between the two measures in these areas. The basis of the FFT is constructivist learning theory and the instrument focuses observers to look for evidence in student-teacher interactions and students' behavior and responses. In contrast, the MyiLOGS IP index focuses on the amount of time teachers use evidence-based practices. Although many of these practices can be considered universal classroom teaching practices, they are often teacher-directed, which is the opposite of the constructivist teaching methodology. This focus on constructivism and students offers a potential explanation for the lack of relationship between MyiLOGS IP and the FFT domains and components. Given that we did not see a negative correlation between the MyiLOGS IP index and scores and the FFT domain and component scores, we can assume that some teachers scoring high on constructivist teaching methods (i.e., the FFT) still make use of universal classroom teaching practices.

The relationship between scores on the CSAS-O and MyiLOGS was similar to that of the CSAS-O and FFT. As CSAS-O discrepancy scores increased, indicating a greater need for change in classroom practices, scores on the MyiLOGS indices decreased.

Specifically, as instruction on the CSAS-O evidenced greater need for change, the MyiLOGS CP index Running Head: Multi-Measure Classroom Observation decreased. The current study's results suggest less effective instruction as indicated by greater need for change parallels less time spent on utilizing teaching strategies focused on higher order critical thinking processes. The current study also found significant negative correlations between the CSAS-O BMS and the MyiLOGS CP, IP, and GF indices, although no parallel relationships were found between the CSAS-O IS and the MyiLOGS IP and GF indices. The relationships between the CSAS BMS rating scale and the MyiLOGS indices provide further support for the MyiLOGS indices, which should share a negative correlation with the CSAS-O. Less effective classroom behavioral management strategy usage parallels a decreases in evidence based teaching strategies and a decrease in focusing on students' higher order critical thinking processes. The finding of no relationship between the CSAS IS Rating Scale and the MyiLOGS IP and GF was unanticipated due to the theoretical content-validity alignment of the MyiLOGS indices and the CSAS-O IS Rating scales descriptions. Both measures emphasize teaching practices related to explicit instruction models. This finding may be due to the aforementioned time-sampling method of MyiLOGS, which records measurement at a different interval than the CSAS-O.

4.3 Implications for Teacher Practice, Professional Development, and Evaluation

Findings from this article offer several trends that have implications for teachers' classroom practice and the school personnel charged with evaluating and supporting the development of teachers. First, we would like to underscore the importance of teachers' classroom behavioral management skills. For the FFT and CSAS-O instruments, the classroom behavior management based scores possessed the most frequent and largest significant negative relationships with both the behavioral and instructional constructs of each measure. This trend was clear when comparing the CSAS-O Strategy Rating scale discrepancy scores to the FFT Running Head: Multi-Measure Classroom Observation scores; most notably, the relationship of the CSAS-O behavioral management constructs to the FFT's instructionally based constructs. As the need for change in classroom behavioral management increased (i.e., larger discrepancy scores) on the CSAS-O, instructional effectiveness on the FFT decreased.

A similar pattern was found for both measures relationship with the MyiLOGS scores. As the use of ineffective and reactive behavioral strategies increased on the CSAS-O, teachers spent less time using the more desirable teaching behaviors measured by the MyiLOGS CP and IP indices (i.e., higher order cognitive processes and evidence based practices). Similarly, as the need for change measured by the CSAS-O Rating scale discrepancy scores increased, the MyiLOGS CP and IP indices decreased. The FFT evidenced significant positive relationships with the MyiLOGS CP index and scores, suggesting that as the classroom environment became more effective on the FFT, teachers' spent more time devoted to students' higher order cognitive processes.

In sum, as classroom behavioral management effectiveness decreased on the CSAS-O and FFT, instructional effectiveness decreased on the CSAS-O, FFT, and MyiLOGS. This underscores the important relation between instruction and classroom management. Specifically, teachers' ability to target students' higher order cognitive processes, which are the focus of the FFT's theoretical orientation and the MyiLOGS CP index. These findings highlight the important impact classroom behavioral management skills have on instructional practice. Simply making instruction more engaging is not guaranteed to produce effective instruction. Classroom behavioral management is an essential skill teachers need to have in order to promote an effective learning environment, especially an environment that aims to focus on higher order critical thinking processes and allows students to be self-directed learners. Running Head: Multi-Measure Classroom Observation Second, in terms of professional development and evaluation contexts, the assessment instrument being used to measure teachers' classroom practices can influence intended outcomes. In the current study, we did not find the FFT, CSAS, and MyiLOGS to converge on all conceptually similar constructs. This suggests that even within construct, measures can differ in their interpretation and even highly reliable and valid observational assessments can present with a specific orientation bias. With this in mind, it is important to remember that teaching is tied to the context in which instruction occurs (Jin&Cortazzi, 1998; Jones & Brownell, 2014; Pratt et. al, 1999) and reliance on any single observational assessment limits the ability to make inferences about effective teaching (Kettler et al., 2017). In the context of special education settings, this understanding about the limitations of using a single observational assessment becomes even more important. Instruction in special education settings typically will appear to be teacher directed (explicit instruction) with a focus on essential concepts, strategies, and skills (Jones, & Brownell, 2014; Brownell, et al., 2012).

Although this may be markedly different than what is considered effective instruction in general education settings according to some teaching models and perspectives, however there is a strong research base supporting this approach to special education contexts (Jones & Brownell, 2014; Brownell et al., 2012).

4.5 Study Limitations

The current study presents with several limitations. Primarily, the generalizability of findings is limited by the size of the sample in regards to teachers and observers. Although a strong component of this study is the use of multiple raters using the same observational instruments to observe the same teachers, the teacher sample only included 10 classroom videos (i.e., 10 unique teachers) observed by 9 unique observers. As such there was not equal representation across grade level or content areas, as well general education and special Running Head: Multi-Measure Classroom Observation education settings for the teachers. Although the observers had backgrounds as teachers and school administrators, their participation in the SSI Project afforded them the opportunity to receive the unique training required for this study, which may not reflect the reality of school-based observers' competencies. Additionally, the demographic characteristics of teachers and observers in the current study may not be representative of all school contexts, particularly high-poverty and rural contexts. Replication with more observers and videos or actual classroom performance would be worthwhile.

A secondary limitation to the current study is that the MyiLOGS observer form was used as a proxy for the MyiLOGS online logging system. MyiLOGS main purpose is a teacher self-report log that tracks teachers' implementation of the three key opportunity to learn indices of time, content, and coverage. We did not include the corresponding teacher log data from MyiLOGS in the current study. As a result, our interpretations of teachers' behaviors according to the MyiLOGS observer form may be limited without the available teacher logs. Future studies should enlist teachers to complete MyiLOGS while observations with the CSAS-O, FFT, and MyiLOGS observer form are conducted.

Lastly, in regards to the MyiLOGS assessment, the time-sampling procedure of the observer form presents with a different unit of measurement than the CSAS-O and the FFT. Whereas the CSAS-O and the FFT look across the entire observed lesson, the MyiLOGS observer form records the dominant cognitive process, instructional practice, and grouping format for each minute of an observed lesson. Thus, the MyiLOGS observer form is potentially limited in its ability capture information related to the concurrent use of multiple teaching practices.

4.6 Future Directions and Research Running Head: Multi-Measure Classroom Observation Research on the concurrent use of multiple classroom observational assessments and the multi-dimensionality of teaching practices is very limited (Ko, Sammons, &Bakkum, 2013). Future research should consider the application of multiple classroom observational assessments across multiple instructional contexts such as grade level, content area, and special education versus general education settings. The outcomes of such research can further our understanding of which teaching practices are effective under specific instructional contexts, thereby leading to better methods of instruction for students, evaluation of effective teaching, and individualized targeted professional development for teachers. With the current study in mind, future research should also consider how the application of multiple classroom observational assessments can predict student achievement, behavioral, and social-emotional learning outcomes. Research may find that the combination of certain measures are better predictors than others and that certain instruments may be better suited for targeted interventions in the classroom to promote specific student needs.

4.7 Conclusion

This article offers a demonstration of the concurrent use of multiple classroom observational assessments to inform teachers' classroom practice. In sum, the CSAS and FFT provide scores in ranges that are related but non-overlapping, and the MyiLOGS Cognitive Processes (CP) scores also appear to be related, but non-overlapping with CSAS and FFT scores. MyiLOGS Instructional Practices (IP) scores appear unrelated to most other scores, which is unexpected because each of these scores should be somewhat related to a general trait of effective teaching. Collectively, these results indicate that many of the scores from the three measures are internally consistent and may be used together to provide non-redundant information about educator performance. Running Head: Multi-Measure Classroom Observation

4.8 Compliance with Ethical Standards

Conflict of Interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent: Informed consent was obtained from all individual participants included in the study.

References

Alberto, P. A., & Troutman, A. C. (2006). Applied behavior analysis for teachers. Upper Saddle River, NJ: Pearson.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R.,Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York, NY: Longman.

Brophy, J. & Good, T. (1986) 'Teacher behavior and student achievement'. In M. Wittrock (ed.), *Handbook of research on teaching* (3rd edn, 328-375). New York: Macmillan.

Danielson, C. (2013). *The framework for teaching evaluation instrument*. Princeton, NJ: The Danielson Group.

Danielson, C. (1996). Enhancing professional practice: A frame-work for teaching (1st ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2007). Enhancing professional practice: A frame-work for teaching (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development. Running Head: Multi-Measure Classroom Observation

Darling-Hammond, L., Herman, J., Pellegrino, J. et al. (2013). Criteria for high-quality assessments. Stanford,CA: Stanford Center for Opportunity Policy in Education, published with the Center for Research on Evaluation, Student Standards, and Testing (CRESST), UCLA, and the Learning Sciences Research Institute, University of Illinois at Chicago.

Dwyer, C. A. (1994). Criteria for performance-based teacher assessments: Validity, standards, and issues. Journal of Personnel Evaluation in Education, 8, 135-150.

Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education*, 79, 79–107.

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., &Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, 79, 1202–1242.

Goe, L., Bell, C., & Little, O. (2008). Approaches to evaluating teacher effectiveness: A research synthesis. Washington, DC: National Comprehensive Center for Teacher Quality.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, 17, 207–219.

Hattie, J., Biggs, J. & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. Review of Educational Research, 66, 99-136.

Horner, R. H., Sugai, G., Todd, A. W., & Lewis-Palmer, T. (2005). School-wide positive behavior support: An alternative approach to discipline in schools. In L. Bambara & L. Kern (Eds.). Individualized supports for students with problem behavior: Designing positive behavior plans. (pp. 359-390), New York: Guilford Press. Running Head: Multi-Measure Classroom Observation

Jones, N. D., & Brownell, M. T. (2014). Assessing the viability of using classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention, 39,* 112-124.

Kane, T., Kerr, K., &Pianta, R. (Eds.). (2014). Designing teacher evaluation systems: New guidance from the measures of effective teaching project. San Francisco, CA: John Wiley.

Kettler, R.J., Arnold-Berkovits, I., Reddy, L.A., Kurz, A., Dudek, C.M., Hua, A., & Lekwa, A. (under review). Use of multi-method and informant teacher evaluation for high poverty schools. *Studies of Educational Evaluation*

Kettler, R.J., & Reddy, L.A. (in press). Using observational assessment to inform professional development decisions: Alternative scoring for the Danielson Framework for Teaching Assessment for Effective Intervention.

Kimball, S., White, B., Milanowski, A., & Borman, G. (2004). Examining the relationship between teacher evaluation and stu-dent assessment results in Washoe County. *Peabody Journal of Education*, 79, 54–78.

Ko, J., Sammons, P., &Bakkum, L. (2013). Effective teaching: A review of research and evidence. Hong Kong: Hong Kong Institute of Education; England: CfBT Education Trust.

Kounin, J. S. (1970). *Discipline and Group Management in Classrooms*. New York: Holt, Rinehart & Winston.

Kurz, A. (2011). Access to what should be taught and will be tested: Students' opportunity to learn the intended curriculum. In S. N. Elliott, R. J. Kettler, P. A. Beddow& A. Kurz (Eds.), Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy (pp. 99–129). New York, NY: Springer. Running Head: Multi-Measure Classroom Observation

Kurz, A., Elliot, S. N., &Shrago, J.S. (2009). MyiLOGS: *My instructional learning opportunities guidance system*. Nashville, TN: Vanderbilt University.

Kurz, A., Elliott, S. N., Kettler, R. J., &Yel, N. (2014). Assessing students' opportunity to learn the intended curriculum: Initial validity evidence for an online teacher log. *Educational Assessment*, 19, 159–184.

Kurz, A., Elliott, S. N., Lemons, C. J., Zigmond, N., &Kloo, A. (2014). Opportunity to learn: A differentiated opportunity structure for students with disabilities in general education classrooms. *Assessment for Effective Intervention.* 40, 1, 24-39. doi: 10.1177/1534508414522685

Jin, L. &Cortazzi, M. (1998) 'Dimensions of dialogue: large classes in China'. *International Journal of Educational Research*, 29, 739-761.

Jones, N. D., & Brownell, M. T. (2014). Assessing the viability of using classroom observations in the evaluation of special education teachers. *Assessment for Effective Intervention, 39,* 112-124.

Marano, R., Pickering, D., & Pollock, J. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement.* Alexandria, VA: Association for Supervision and Curriculum Development.

Mead, Sara (2012). "Recent State Action on Teacher Effectiveness: What's in State Laws and Regulations?". Washington, DC: Bellwether Education Partners.

Milanowski, A. T. (2011). Validity research on teacher evaluation systems based on the Framework for Teaching. Madison, WI: Consortium for Policy Research in Education. Running Head: Multi-Measure Classroom Observation

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, 79(4), 33-53.

Milanowski, A., Kimball, S.M., & White, B. (2004). *The relationship between standards-based teacher evaluation scores and student achievement.* Madison, WI: University of Wisconsin-Madison, Consortium for Policy Research in Education.

Muijs, D., & Reynolds, D. (2005) Effective teaching: evidence and practice (2nd edn). London: Sage.

Nelson, P., Reddy, L.A., Dudek, C.M., & Lekwa, A. (2017). Observer and Student Ratings of the Class Environment: A preliminary investigation of convergence. *School Psychology Quarterly.* 31 (3), 1-15.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom Assessment Scoring System: Manual K-3. Baltimore, MD: Paul J. Brookes.

Peterson, L. (1995). Stop and think learning: A teacher's guide for motivating children to learn: including those with special needs. Camberwell, Victoria: The Australian Council for Educational Research.

Porter, A. C., Youngs, P., and Odden, A. (2001). Advances in teacher assessments and their uses. In V. Richardson (Ed.), Handbook of Research on Teaching, Fourth Edition. Washington, DC: American Educational Research Association, 259-297.

Pratt, D., Kelly, M. & Wong, W. (1999) 'Chinese conceptions of effective teaching in Hong Kong: toward culturally sensitive evaluation of teaching'. *International journal of lifelong education*, 18(4), 241-258. Running Head: Multi-Measure Classroom Observation

Reddy, L. A., & Dudek, C. M. (2014). Teacher progress monitoring of instructional and behavioral management practices: An evidence-based approach to improving classroom practices. *International Journal of Education and Psychology*, *2*, 71-84.

Reddy, L. A., Dudek, C. M., Fabiano, G. A., & Peters, S. (2015). Measuring teacher self-report on classroom practices: Construct validity and reliability of the Classroom Strategies Scale-Teacher Form. *School Psychology Quarterly*, 30, 513-533.

Reddy, L.A., Fabiano, G. A., Dudek, C. M. (2013). Concurrent validity of the Classroom Strategies Scale for elementary school – Observer form. *Journal of Psychoeducational Assessment, 31,* 258-270.

Reddy, L. A., Fabiano, G. A., Dudek, C. M., & Hsu, L. (2013a). Development and construct validity of the Classroom Strategies Scale– Observer Form. *School Psychology Quarterly, 28, 317-341*.

Reddy, L. A., Fabiano, G. A., & Dudek C. M. &. Hsu, L. (2013b). Predictive validity of the Classroom Strategies Scale-Observer form on statewide testing scores: A preliminary investigation. *School Psychology Quarterly, 28,* 301-316.

Rowe, K. (2006) Effective teaching practices for students with and without learning difficulties: Issues and implications. *Australian Journal of Learning Disabilities*, 11(3), 99-115.

Scheerens, K. (1992) *Effective schooling: research, theory and practice*. London: Cassell.

Stage, S. A., Quiroz, D. R., (1997). A meta-analysis of interventions to decrease disruptive classroom behavior public education settings. *School Psychology Review*, 26 (3), 333-368.

Steele, J. M., House, E. R., &Kerins, T. An instrument for assessing instructional climate through low-inference student judgements. *American Educational Research Journal*, 1971, 8, 447-466. Running Head: Multi-Measure Classroom Observation

Vaughn, S., Gersten, R., & Chard, D. J. (2000). The underlying message in LD intervention research: Findings from research syntheses. *Exceptional Children,* 67, 99–114.

Walberg, H. J. (1986) 'Syntheses of research on teaching'. In M. Wittrock (ed.), *Handbook of research on teaching* (3rd edn, pp. 570-602). New York: Macmillan.

Walker, H. M., Ramsey, E., & Gresham, F. M. (2003). Heading off disruptive behavior: How early intervention can reduce defiant behavior and win back teaching time. *American Educator,* 26, (4), 6-45.

Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 155–180). Mahwah, NJ: Lawrence Erlbaum.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. Brooklyn, NY: The New Teacher Project. Running Head: Multi-Measure Classroom Observation